Using Statistical and Machine Learning Approaches to Predict Study-dropout in ALSPAC

Abstract:

Background:

The Avon Longitudinal Study of Parents and Children (ALSPAC) offer invaluable insights for understanding child development from birth to adulthood. However, like many longitudinal studies, it suffers from participant dropout and missing data. Our analysis employs statistical and machine learning methods to predict dropout rates and identify influencing factors to improve the study's reliability and future interventions.

Methods:

Our study employs Random Forest models and ten-fold cross-validation to investigate participant dropout at age 11 within the ALSPAC dataset. Using a Gini decrease measure, we identified and categorised the top 10% impactful variables into six domains. The analysis incorporates logistic regression and correlation tests to understand the factors affecting dropout.

Results:

The Random Forest model achieved mean AUCs of 0.75 for questionnaire data and 0.74 for clinic data in predicting participant dropout at age 11 within the ALSPAC dataset. Key predictors were mainly in the Demographic and Lifestyle domains. A Pearson correlation of 0.98 between clinic and questionnaire data supports the consistency of these predictors.

Discussion:

The findings extend previous research by considering various impacting factors and suggest that targeted interventions could improve retention in longitudinal studies. The study's robust methodology and alignment across data collection methods reinforce its validity, although its focus on a single cohort limits generalizability.

Keywords: ALSPAC, dropout prediction, Random Forest, Gini Decrease, participant retention, longitudinal research

Introduction:

Longitudinal cohort studies, such as the ALSPAC, have significantly contributed to understanding how children grow and develop from birth to adulthood [3]. This helps us understand the associations between health, behaviour, and development over time. However, a challenge in these studies is the ongoing problem of participants dropping out or missing data. The phenomenon of missing data can substantially undermine a study's statistical integrity, leading to biased outcomes. Since the remaining subjects may not represent a random subset of the original population, this poses serious questions regarding the authenticity and reliability of the findings, amplifying the potential for incorrect interpretations and applications [26]. Missing data can compromise the study's internal validity, affecting the accuracy of conclusions within the study and its external validity, limiting the generalizability of findings. Selection bias further complicates this, where the remaining participants may need to be more representative, skewing results and conclusions.

The ALSPAC study, which has been a vital resource for academic research, offers an opportunity to investigate further the factors contributing to dropout at different stages of development [27]. This gap encompasses the longitudinal assessment of dropout, identifying specific cultural and socioeconomic indicators, the evaluation of targeted intervention effectiveness, and the identification of distinct predictors associated with various stages of life.

The main aim of this study is to address existing knowledge gaps about participant dropout within the ALSPAC dataset, with a particular focus on the age of 11. Our study has two main objectives: to predict the likelihood of missing data at this age and to identify the participant characteristics that influence dropout. Utilising machine learning and statistical methods, the primary motivation for this study is the necessity to perform a comprehensive and impartial investigation of many factors, both anticipated and unanticipated, that contribute to the discontinuation of studies. Despite its considerable significance for longitudinal research, this particular area of inquiry has received limited attention in the existing literature [1]. This study seeks to contribute to the current body of knowledge and enhance the effectiveness of treatment interventions.

Methods:

The methods section outlines this research study's various processes and strategies, focusing on participant dropout. This section is organised into six main sub-headings, each designed to provide detailed insights into the study's design and implementation.

Data Source:

The data for this study is sourced from the Avon Longitudinal Study of Parents and Children (ALSPAC), a UK cohort study on developmental trajectories [3,13]. Initially, ALSPAC enrolled 14,541 pregnant women between April 1991 and December 1992, leading to 14,062 live births and 13,988 children alive in one year. Additional recruitment phases added 906 new pregnancies, bringing the sample to 15,447 pregnancies and 15,658 fetuses, with 14,901 children surviving the first year [3].

While the ALSPAC data is expansive, this study narrows its focus to a core sample from the original enrolment, comprising 14,440 records of mothers and their corresponding children. This focused subsample allows us to investigate the predictors of dropout at age 11 with data that combines mother-child information, providing a complete perspective.

Please note that the study website contains details about all available data through a fully searchable data dictionary and variable search tool.

Data in ALSPAC are collected through a variety of methods, including:

- **Childhood Stages**: Observations and measurements taken at various developmental stages of the children's lives at clinics.
- **Parental Questionnaires**: Surveys filled out by the mothers provide information about the child and the home environment.

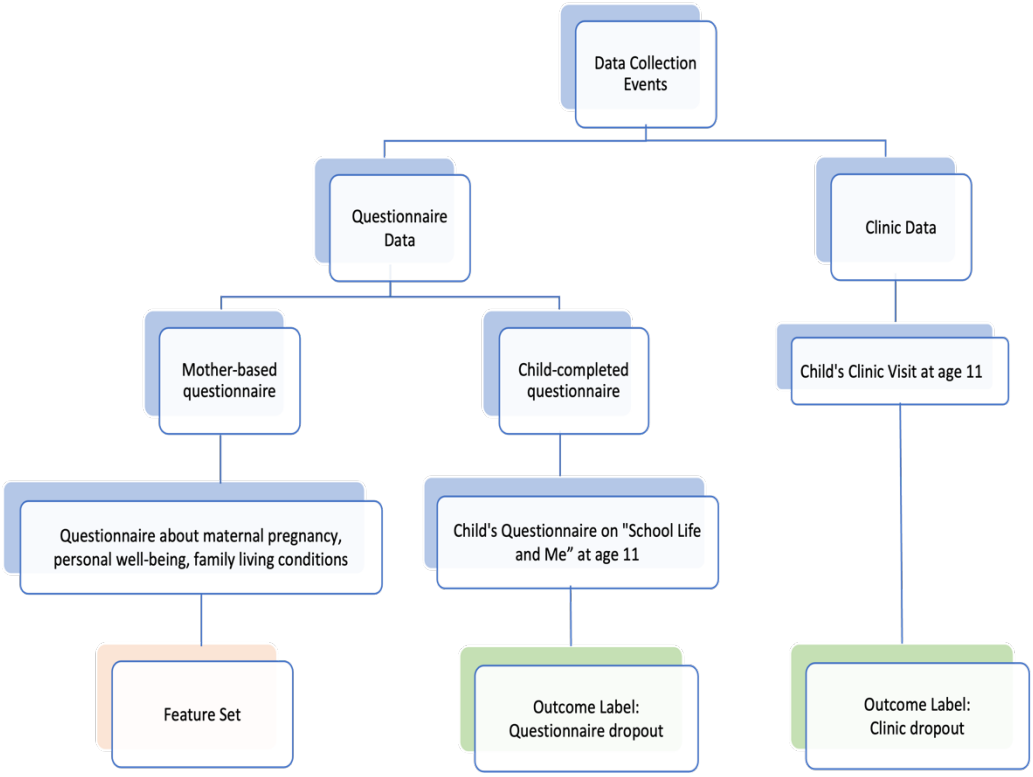

Figure 1: Data Collection and Utilization Overview

Figure 1 illustrates the data sources used in our study and their specific applications. Two primary categories of data, Mother Questionnaire Data and Child Data (both questionnaire and clinic) are outlined.

The choice of ALSPAC data was motivated by its ethical rigour and the breadth of variables it contains, which allows for an in-depth analysis of dropout rates without bias.

Software and Tools:

The analysis in this study was performed using R (version 4.2.1) [23]. The Caret package was utilised for modelling purposes [18], while ggplot2 was employed for data visualisation [28]. Packages such as Random Forest [20] and Random Forest Explainer played a crucial role in predictive modelling, while Git provided essential functionalities for version control and creating reproducible environments [6]. The code used for this research is available in the GitHub repository at https://github.com/MRCIEU/predicting-ALSPAC-dropout.

Variable Selection:

In our hypothesis-free study, a comprehensive set of variables was selected to act as predictors. These variables were sourced from diverse data collection methods, such as mother and child

questionnaires and clinical visits at age 11. After a systematic data cleaning and processing pipeline, these predictors cover various factors, from environmental conditions to individual health and behavioural patterns [3, 13].

The outcome variables, representing the level of participation in different data collection events throughout childhood, were created as 'questionnaire dropout' and 'clinic dropout' indicators [24]. Figure 1, a detailed timeline diagram, further illustrates the selection and processing of these variables. This meticulous approach enhances our understanding of subjects' experiences and involvement across various developmental stages.

Data Cleaning and Pre-processing:

Ensuring the data's quality, consistency, and accuracy was crucial, given that the ALSPAC dataset is large and multi-dimensional, affecting various aspects of human development. Any inaccuracies or inconsistencies could undermine the validity of the research findings [13]. This stage involved a series of systematic choices and steps to manage the dataset effectively. These steps were designed to provide a thorough understanding of the variables in the dataset and their specific characteristics [3]. The procedures included:

A. Classification of Variables:

Our study classified variables using a manual approach inspired by the PHESANT (Phenome-Scan Analysis Tool) methodology [22]. Each variable was analysed to determine its type—ordinal, nominal, binary, or continuous—based on inherent characteristics [16]. Nominal variables were further converted to binary types where suitable. Additional details and code are available in the Supplementary Material and the accompanying GitHub repository.

B. Handling Negative Values:

Upon detailed review of the dataset, we discovered negative values in specific variables, such as "age" and "BMI," that logically cannot be negative. After a comprehensive analysis, these inconsistencies were categorised as missing data to maintain analytical integrity [19]. This approach ensures that our interpretation remains unbiased.

C. Imputation of Missing Values:

In the data analysis, missing values were identified and treated using an imputation strategy. For continuous variables, the missing values were replaced with the median, while for categorical variables, the mode was used. This approach is a standard feature in the Random Forest machine learning package [20]. The method helps to maintain the central tendency of the data distribution and minimises the introduction of bias, making the statistical analysis more robust.

D. Outlier Analysis:

We first applied automated statistical tests on each variable to identify outliers, flagging values that fell outside 1.5 times the interquartile range (IQR) as potential outliers [2]. This initial step was followed by manual verification through graphical methods like histograms. For instance, a data point indicating an age of '140 years' was flagged and removed due to its implausibility. After this validation, we confirmed that no significant anomalies were left in the dataset, thereby assuring its integrity for further analysis.

E. Ensuring Data Quality:

The last step in data cleaning was validating the entire pre-processing workflow. Quality checks were conducted to ensure the dataset's integrity, consistency, and reliability. The tasks included verifying data completeness, checking value consistency, conducting range validation, removing duplicate records, and performing cross-field validation. Inconsistencies and distortions were removed by checking the dataset, making it ready for precise and reliable analysis.

Our study followed a thorough data cleaning and pre-processing protocol to prepare the dataset for analysis. Variables were classified to align with analytical techniques, and negative values were appropriately managed. Missing values were imputed to accommodate the limitations of the Random Forest algorithm, while outlier analysis confirmed the data's reliability. Together, these measures created a reliable dataset that supports accurate research outcomes.

Statistical and Machine Learning Methods:

The methodology includes a focus on optimising predictor selection. Zero-variance predictors were identified when all their values remained constant across all observations. These

153 predictors were eliminated as they hold no predictive power [16]. Variable selection was
154 further refined by evaluating for near-zero variance [8], which refers to predictors that almost,
155 but not entirely, have a single constant value [17]. This was undertaken to improve the model's
156 performance and computational efficiency.

157 Random forests were selected due to their efficacy in handling large datasets with higher
158 dimensionality [4]. Random forests efficiently deal with potential overfitting and enhance
159 generalisation through bootstrapped aggregating and utilising numerous decision trees [5,9].
160 The "mtry" parameter in Random Forest models determines the number of characteristics
161 considered while searching for the optimal split at each node in the tree. Using randomness in
162 the model enhances generalisation capabilities [14]. The adjustment of the mtry parameter may
163 have a substantial influence on the performance of the model. If the value of mtry is higher,
164 there is a risk of overlooking crucial characteristics. Conversely, if the value is too high, the
165 model may become less varied and more susceptible to overfitting. The use of random forests
166 is backed by academic literature and empirical evidence demonstrating their utility in complex
167 predictive modelling, especially in scenarios where the relationships between predictors and
168 outcomes are non-linear [4].

169 Our study integrates a 10-fold cross-validation scheme within a Random Forest model to
170 optimise the Receiver Operating Characteristic (ROC) curve [24]. It is important to note that
171 while this approach is comprehensive, extending beyond standard cross-validation by
172 incorporating additional steps such as storing final predictions and computing class
173 probabilities, it does not engage in specialised hyperparameter tuning [4,10]. The rationale
174 behind employing 10-fold cross-validation lies in its ability to balance computational efficiency
175 with reliable performance assessment judiciously. Although the focus is on the AUC (Area
176 Under the Curve) and its corresponding ROC value, we also acknowledge the utility of
177 supplementary metrics, particularly precision and recall, especially in situations with an
178 imbalanced class distribution [9].

179 For a more precise visualisation of our methodological approach, Figure 2 presents a detailed
180 flowchart outlining each step, from data pre-processing to model evaluation. Additional
181 explanations are provided in the supplementary material.

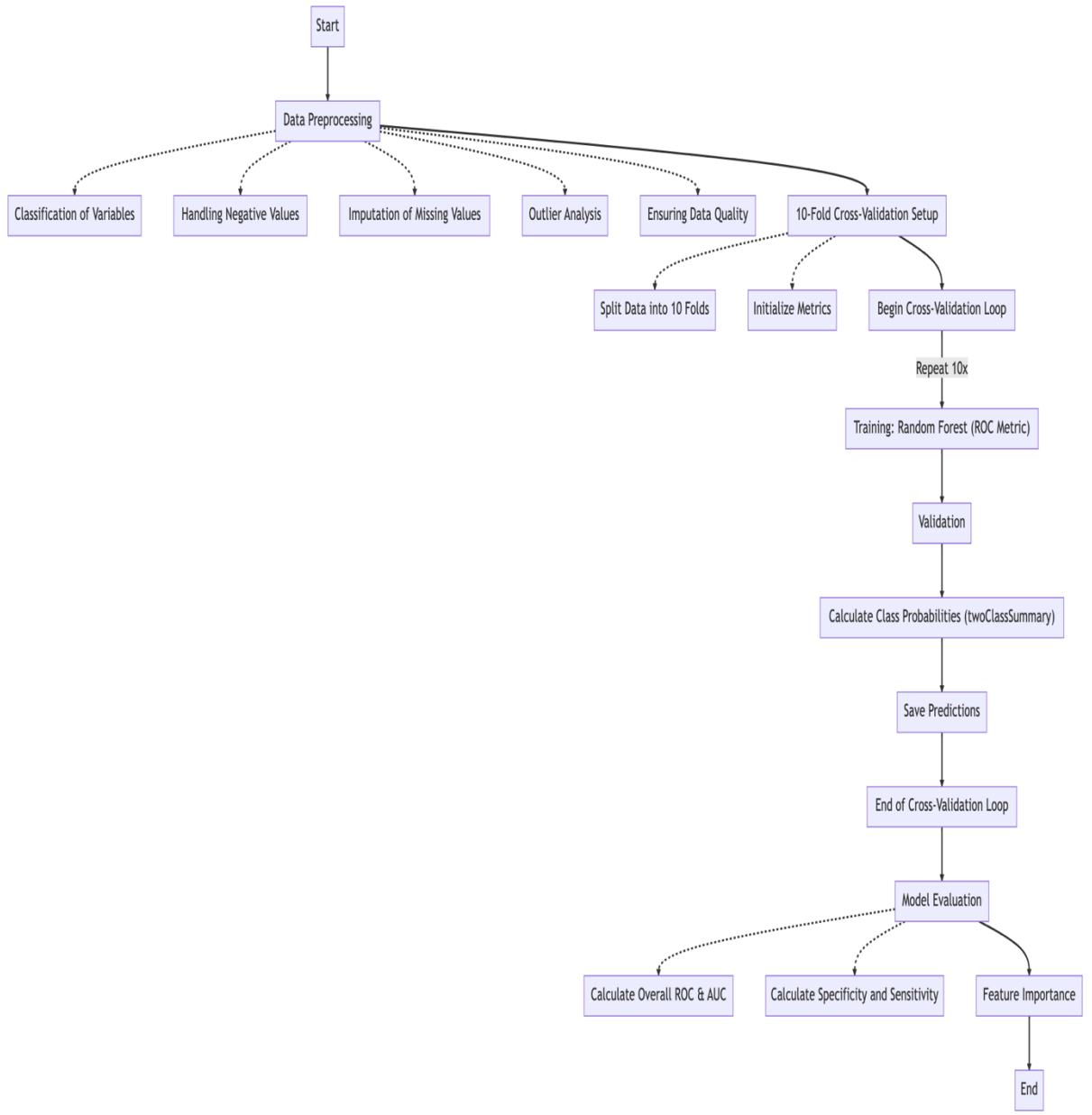

Figure 2: Detailed flowchart outlining the steps from data pre-processing to model evaluation

The Gini decrease measure was applied within Random Forest models to quantify each predictor's importance in the dropout rates of clinics and questionnaires [4]. The Gini decrease helps identify each feature's importance by calculating its contribution to making the target variable more uniform. A higher Gini decrease indicates a more critical role in classifying the target variable [12]; more details about this are available in the supplementary section. Instead of a simultaneous assessment, both models were systematically compared to uncover correlations and meaningful patterns among the predictors.

Using Gini decrease values, we identified the top 10% most impactful variables, which led to a subset of 206 variables. After identifying these top 206 variables, we sorted them into six specific categories based on what each variable measured. The categories were not predetermined; instead, we created them after we knew which variables were most impactful.

The domains identified are as follows:

1. **Demographic Information:** Includes factors like ethnicity, socioeconomic status, and location.
2. **Lifestyle Factors:** Consists of diet, exercise, and substance use variables.
3. **Health Metrics:** Covers medical history, physical exam results, and similar health-related variables.
4. **Psychological Measures:** Includes variables related to mental and emotional well-being and personality traits.
5. **Social Factors:** Focuses on social relationships, support systems, and community involvement.
6. **Pregnancy and Child Care:** Includes maternal health, prenatal care, and early childcare variables.

By categorising variables this way, we aim for a more focused analysis of the predictors that impact participant dropout at age 11. To verify these relationships, we relied on established statistical methods alongside visual aids like scatter plots and bar graphs for initial examinations.

We used the GLM (Generalised Linear Model) function in R to perform logistic regression on clinic and questionnaire datasets, focusing specifically on predictors with a Gini decrease value above 40, which aims to identify the directional influence of specific predictors of dropout.

Pearson's correlation test was explicitly employed to evaluate the linear relationship between changes in the Gini coefficient across clinic and questionnaire data [12]. We also conducted a paired t-test to examine if the observed disparities in Gini decrease between the two settings were meaningful. This approach ensures a clear analysis, highlighting key factors affecting dropout rates.

This study employs statistical precision, data visualisation, and predictive modelling techniques to deepen our understanding of the research question. The dropout prediction methodology utilises algorithms like random forest, thorough pre-processing, validation methods, and an exhaustive strategy to assess dropout at age 11. Various research perspectives align with current predictive modelling approaches and facilitate detailed analysis and significant findings.

Ethical Considerations:

This study adheres to stringent ethical guidelines, ensuring the sanctity of ethical research practices. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees, per the standards outlined by the University of Bristol. Specific details on the ethics committee and institutional review boards that approved aspects of the study can be found on the University of Bristol's webpage for research ethics related to the Avon Longitudinal Study of Parents and Children. Informed consent for data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. The ethical framework encompassed considerations such as confidentiality, participants' informed consent, and data protection compliance, safeguarding the research's integrity and the participants' rights and privacy.

Results:

Sample Information:

The study sample had data from 15,414 pregnancies drawn from the core ALSPAC sample size of 14,509.

Descriptive Statistics:

The average maternal age at delivery was 28.12 years, with a standard deviation of 4.47. The mean gestational age was 22.74 weeks, with a standard deviation 4.07. These statistics are summarised in Table 1.

Table 1: Summary of Cohort Demographics, Maternal and Gestational Ages, and Dropout Rates

	Variable	Value	Mean (SD)	Rate (%)
Overall Cohort Demographics				
	Total Pregnancies Identified	15,414	..	..
	Core ALSPAC Sample Size	14,509	..	..
General Descriptive Statistics				
	Maternal Age at Delivery	..	28.12 (4.47)	..
	Average Gestational Age	..	22.74 (4.07)	..
Survival Rates				
	Clinic Dropout Rate	..	..	53.63%
	Questionnaire Dropout Rate	..	..	48.41%
Comparative Maternal Age Statistics by Dropout Status				
	Clinic: No Dropout	..	22.36 (4.06)	..
	Clinic: Dropout	..	23.07 (4.04)	..
	Questionnaire: No Dropout	..	22.35 (4.08)	..
	Questionnaire: Dropout	..	23.15 (4.01)	..

- **ALSPAC** - Avon Longitudinal Study of Parents and Children.
- **SD** - Standard Deviation.
- **Rate (%)** - Represents the percentage of the cohort falling under the specific category.
- ·· - Indicates cells where data are not applicable.

Survival Rates:

Based on our analysis, the survival rates indicate promising figures. The 28-day survival rate was 95.33%, while the 1-year survival rate slightly had a lower value at 95.19%.

Dropout Rates:

The clinic dropout rate was 53.63%, and the questionnaire dropout rate was 48.41%.

Comparative Statistics for Dropouts:

In terms of comparative statistics for those who dropped out versus those who did not, the data showed some differences in maternal age. These details are also provided in Table 1.

Model Performance:

The Random Forest models were applied to classify dropout patterns within the cohort, encompassing 14,440 samples with 1,502 predictors. The mtry parameter, specifying the number of variables considered for splitting at each tree node, was optimised to a value of 67

for both the questionnaire and clinic dropout models. The classification aimed to identify two classes: 'No Dropout' and 'Dropout'. The models were validated using tenfold cross-validation, ensuring robust assessment across various subsets of the data.

In the tenfold cross-validation for questionnaire dropouts, the Random Forest model with an optimised mtry value of 67 displayed a mean AUC (Area Under the Curve) of 0.75 (95% CI: 0.74–0.76), indicating a robust performance across different data subsets. The standard deviation of the AUC values across the ten folds was 0.016.

For clinic dropout classification, a similarly configured Random Forest model with mtry set to 67 yielded a mean AUC (Area Under the Curve) of 0.74, with a standard deviation of 0.014 across the tenfold cross-validation. The 95% CI for the mean AUC was 0.73–0.75, indicating stable model performance across the different folds of data.

The ROC curves for both questionnaire and clinic dropout are illustrated in Figures 3 and 4, respectively. These graphical representations visually assess the model's classification performance for each scenario.

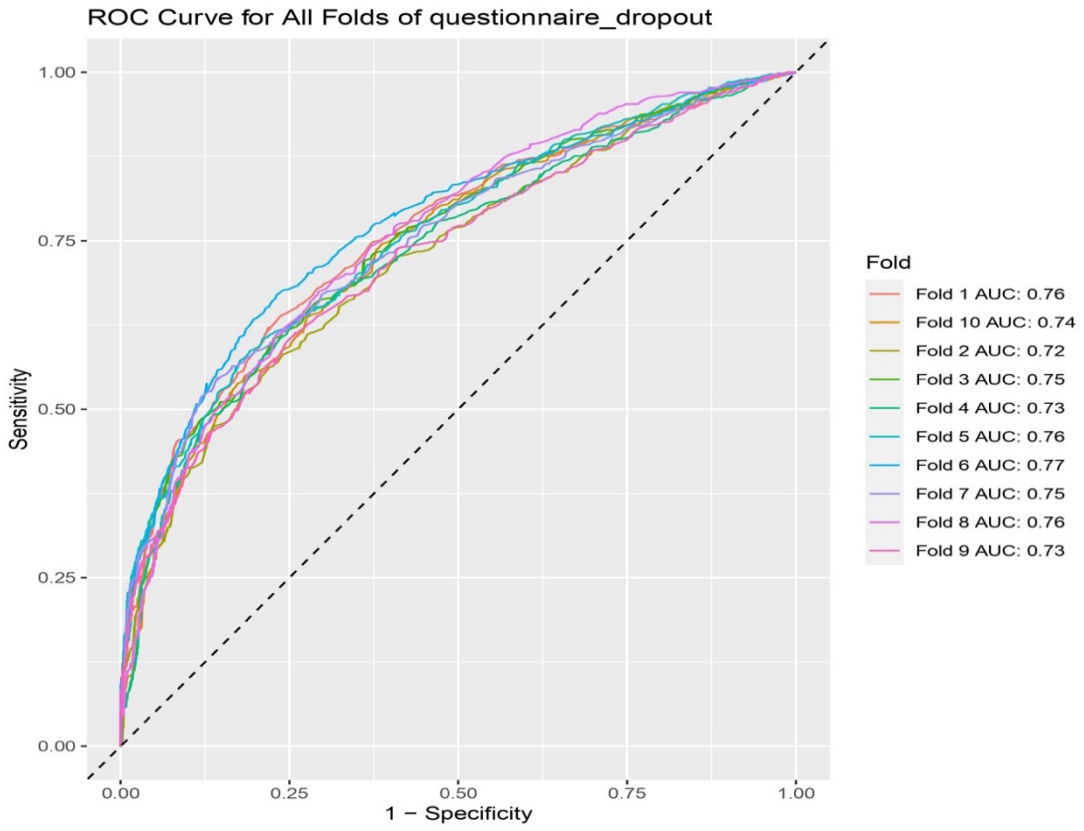

Figure 3: ROC curves representing all ten folds from the tenfold cross-validation for the Random Forest model in the questionnaire dropout scenario.

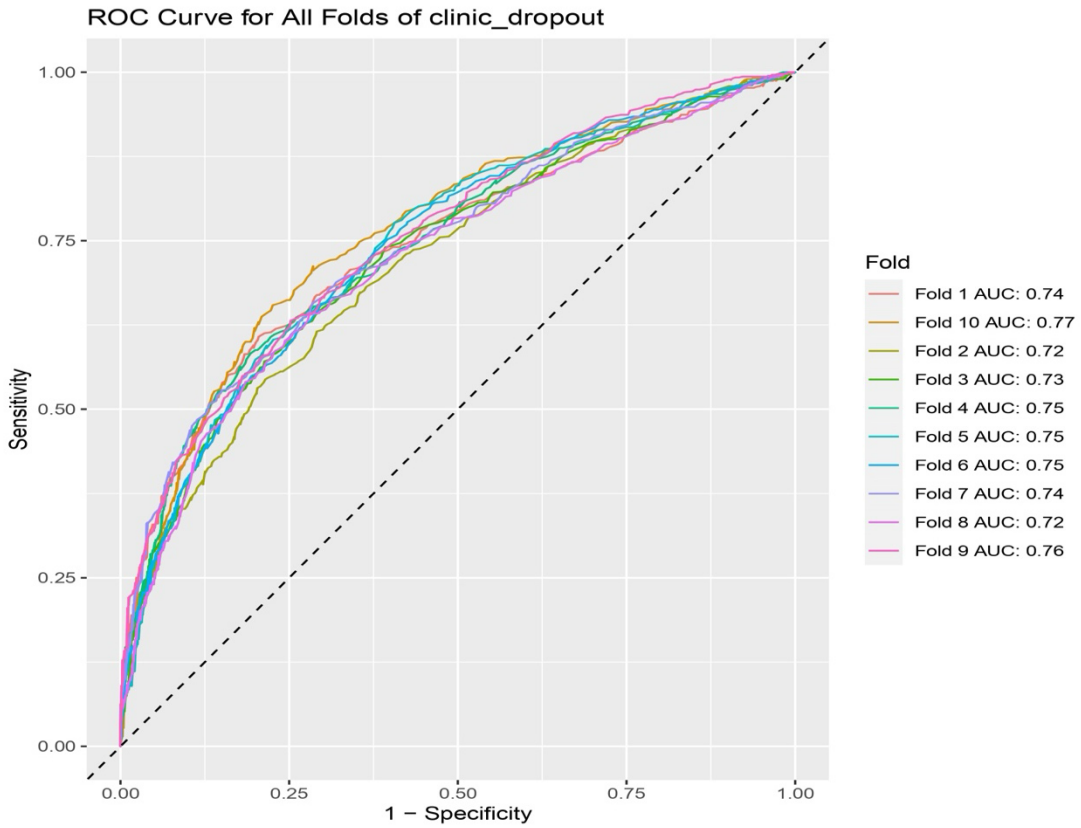

Figure 4: ROC curves representing all ten folds from the tenfold cross-validation for the Random Forest model in the clinic dropout scenario.

Each curve in Figures 3 and 4 depicts a different fold's trade-off between sensitivity and specificity. The diversity of angles demonstrates the model's general performance and variability across different subsets of the data.

Key Predictors:

A comparison bar plot (Figure 5) reveals the distribution of variables both in the initial feature set and among the top 10% most impactful predictors across these six categories.

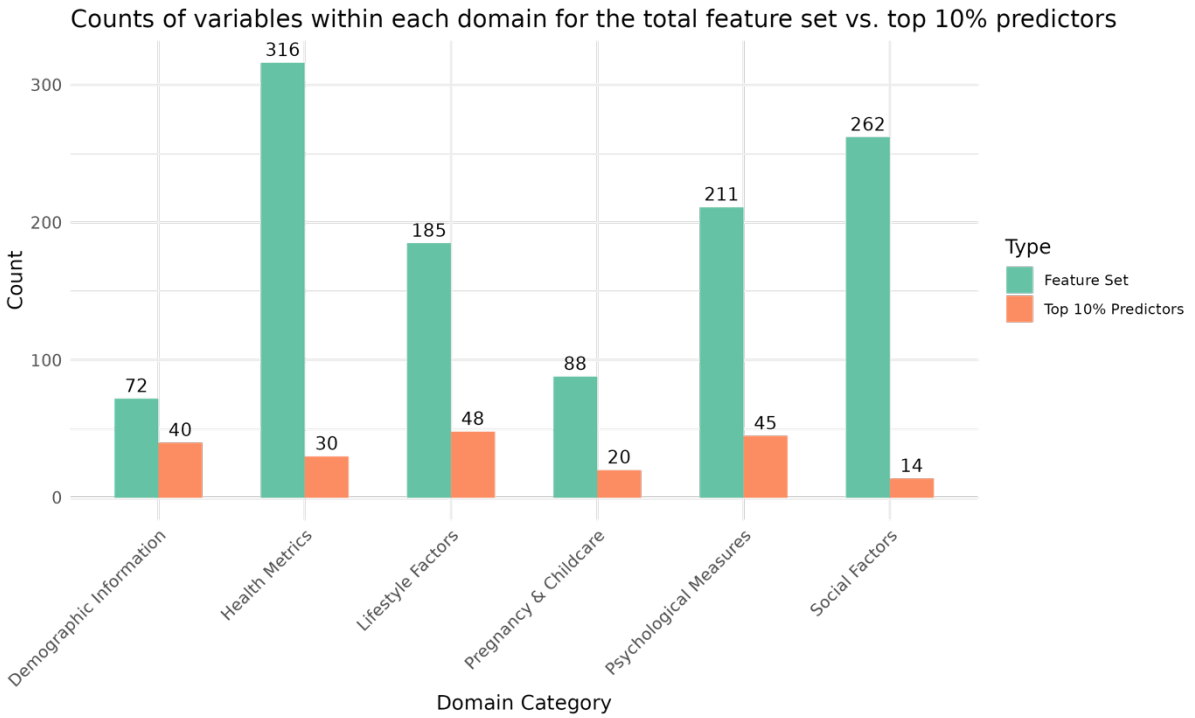

Figure 5: Comparison bar plot showing the distribution of variables (initial feature set vs. top 10%) within the six domains

Figure 6 visually represents the combined importance of the top 10% of features, as measured by their Gini decrease values, for both the clinic and questionnaire data. This plot exhibits a compelling correlation between the clinic and questionnaire data, further emphasising the robustness of these key predictors across different settings.

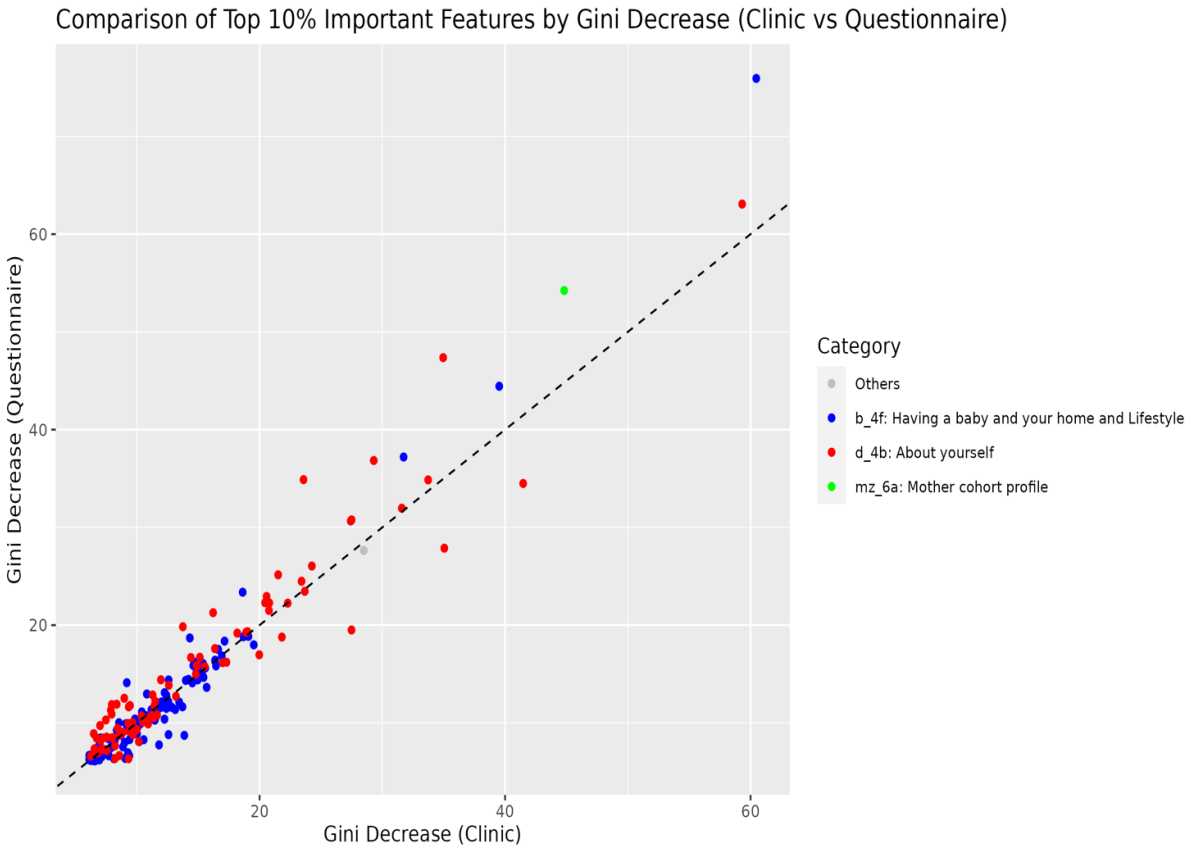

Figure 6: Comparison Scatter plot showing the top 10% important features by Gini decrease (clinic vs. questionnaire)

Logistic regression models were employed to better understand the factors influencing clinic and questionnaire dropout rates. Variables with a Gini importance score greater than 40 were examined for their influence. The table provides a broad summary of these key predictor variables, outlining their Gini scores, coefficients, the direction of their influence on the outcome, and the type of information they represent.

Table 2: Summary of Important Predictor Variables for Clinic and Questionnaire Dropout Models

Dropout Type	Variable	Gini Score	Coefficient	Direction of Influence towards Dropout	Domain Type
Questionnaire	b023	75.94	-0.08	Lowers the risk	Demographic Information
Questionnaire	d994	63.08	0.01	Raises the risk	Demographic Information
Clinic	b023	60.47	-0.06	Lowers the risk	Demographic Information
Clinic	d994	59.31	0.03	Raises the risk	Demographic Information
Questionnaire	mz028b	54.23	-0.14	Lowers the risk	Demographic Information
Questionnaire	d375	47.37	-0.01	Lowers the risk	Demographic Information
Clinic	mz028b	44.82	-0.06	Lowers the risk	Demographic Information
Questionnaire	b925	44.44	0.10	Raises the risk	Demographic Information
Clinic	d2902	41.47	0.87	Raises the risk	Lifestyle Factors

Statistical Analysis:

Pearson's correlation test indicated a remarkably high correlation between the Gini decrease values from the clinic and questionnaire data sets ($r = 0.98$, $p < 2.2e-16$). This high correlation coefficient suggests a strong alignment between the two data collection methods, implying that they are both robust in identifying the same influential variables that impact dropout rates.

In addition to the correlation analysis, to test the mean differences between the Gini decrease values from the clinic and questionnaire models, a paired t-test was performed. The test revealed no considerable mean difference ($t = -0.60$, $df = 2303$, p-value = 0.55). This finding reinforces the idea that both methods consistently identify the predictive factors influencing dropout rates.

Discussion:

The primary focus of this study was to assess how well missingness or dropout at age 11 can be predicted in the ALSPAC cohort and to identify influential predictors that could help reduce dropout in future studies. The Random Forest models provided predictive performance with mean AUCs of 0.75 and 0.74 for the questionnaire and clinic data, respectively. Key predictors primarily fell within the Demographic Information and Lifestyle Factors domain, with notable variables such as maternal age having Gini scores above 60 and opposite directional influences on dropout risk. The Pearson's correlation coefficient of 0.98 between clinic and questionnaire Gini decrease values reinforced the consistency and generalizability of these predictive factors.

In the context of existing research, our study adds another layer to our understanding of participant dropout in longitudinal studies, specifically in the ALSPAC cohort. Previous findings [29] found that dropout in ALSPAC was systematic and associated with disruptive behaviour disorders in children. While they examined the impact of family variables, our research extends this by considering a wide array of factors, including demographic information and lifestyle factors, using advanced machine learning techniques. The high predictive power of our models suggests a more comprehensive approach to understanding the characteristics of those who are more likely to drop out, which could be beneficial in addressing the bias introduced by systematic dropouts.

Another previous finding [27] explored the role of genetic factors in participant dropout and found that various polygenic scores were associated with different participation rates. Our

study, while not focused on genetic markers, reinforces that dropout is not random but influenced by specific identifiable variables. The robustness and consistency in predictors we identified across different data collection methods provide further confidence in targeted intervention strategies. Moreover, our finding that clinic and questionnaire methods yield similar influential variables could serve as a foundation for multi-modal approaches in future longitudinal studies, thus enriching our understanding of the complexities of dropout phenomena.

The strength of this study lies in its methodology, including the use of tenfold cross-validation and the large sample size of 15,414 pregnancies, which adds to the reliability of the predictive model. However, the study has several limitations. It exclusively focuses on the ALSPAC dataset, which may limit the generalizability of findings to other cohorts. The age-specific focus on 11-year-olds further narrows the applicability, as a broader age range covering late childhood or early adulthood might provide more comprehensive insights. Additionally, while the Random Forest model demonstrates strong predictive capabilities, its complexity makes it less interpretable than simpler models like logistic regression.

In summary, our study successfully achieves its objectives by identifying robust predictors of dropout at age 11 within the ALSPAC dataset. The high degree of alignment between the clinic and questionnaire data reaffirms the validity of these predictors across different data collection methods. The findings have meaningful implications for designing targeted interventions to improve cohort retention in longitudinal studies. Future work should validate these predictors in different populations to ascertain their generalizability.

References:

[1] Abshire M, Dinglas VD, Cajita MIA, et al. Participant retention practices in longitudinal clinical research studies with high retention rates. BMC Med Res Methodol. 2017;17:30. doi:10.1186/s12874-017-0310-z.

[2] Aggarwal CC, Yu PS. Outlier detection for high dimensional data. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data; 2001 May 21-24; Santa Barbara, California: 37-46. doi:10.1145/375663.375668.

[3] Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2013;42(1):111-127. doi:10.1093/ije/dys064

[4] Breiman L. Random forests. Machine Learning. 2001;45:5-32. doi:10.1023/A:1010933404324.

[5] Cascarano A, Mur-Petit J, Hernández-González J, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artif Intell Rev. 2023. doi:10.1007/s10462-023-10561-w.

[6] Chacon S, Straub B. Pro Git. 2nd ed.: Apress; 2014.

[7] Chapman P, Clinton J, Kerber R, et al. CRISP-DM 1.0: Step-by-step Data Mining Guide. DaimlerChrysler; 2000.

[8] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc A. 1995;158(3):419-466. doi:10.2307/2983440.

[9] Chen S, Grant E, Wu TT, Bowman FD. Statistical Learning Methods for Longitudinal High-dimensional Data. *Wiley Interdiscip Rev Comput Stat*. 2014;6(1):10-18. doi:10.1002/wics.1282

[10] Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research. Sage Publications; 2017.

[11] Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. 2018;362:k601. Published 2018 Jul 12. doi:10.1136/bmj.k601

[12] DeGroot MH, Schervish MJ. Probability and Statistics. 4th ed. Addison-Wesley; 2012.

[13] Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers' cohort. *International Journal of Epidemiology* 2013; 42:97-110.

[14] Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using Random Forests. Pattern Recognit Lett. 2010;31:2225-2236. doi:10.1016/j.patrec.2010.03.014.

[15] Golding J, Pembrey M, Jones R; ALSPAC Study Team. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol*. 2001;15(1):74-87. doi:10.1046/j.1365-3016.2001.00325.x

[16] Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. 7th ed. Pearson; 2010.

[17] Iglewicz B, Hoaglin D. The ASQC Basic References in Quality Control: Statistical Techniques. In: Mykytka EF, ed. How to Detect and Handle Outliers. Vol 16. ASQC Quality Press; 1993.

[18] Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1-26. doi:10.18637/jss.v028.i05.

[19] Kwak S, Kim J. Statistical data preparation: Management of missing values and outliers. Korean J Anesthesiol. 2017;70:407. doi:10.4097/kjae.2017.70.4.407.

[20] Liaw A, Wiener M. Classification and regression by Randomforest. R News. 2002;2:18-22. Available from: http://CRAN.R-project.org/doc/Rnews/

[21] Little RJ, Rubin DB. Statistical Analysis with Missing Data. Vol 793. John Wiley & Sons; 2019. doi:10.1002/9781119482260.

[22] Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int J Epidemiol*. 2018;47(1):29-35. doi:10.1093/ije/dyx204

[23] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org

[24] Rothman KJ, Greenland S, Lash T. Modern Epidemiology. 3rd ed. Lippincott Williams & Wilkins; 2008:303-327.

[25] Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc; 1987. doi:10.1002/9780470316696.

[26] Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. Published 2009 Jun 29. doi:10.1136/bmj.b2393

[27] Taylor AE, Jones HJ, Sallis H, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2018;47(4):1207-1216. doi:10.1093/ije/dyy060.

[28] Wickham H. Ggplot2: Elegant Graphics for Data Analysis. Springer International Publishing; 2016.

[29] Wolke D, Waylen A, Samara M, et al. Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *Br J Psychiatry*. 2009;195(3):249-256. doi:10.1192/bjp.bp.108.053751

Supplementary Material:

The content within the "Data source," "Ethical considerations," and "Data dictionary" sections has been composed in strict compliance with established guidelines. Specifically, these guidelines are outlined in the ALSPAC Publications Checklist provided by the University of Bristol. The content in these sections is not subject to interpretation or modification and must be presented as mandated by the checklist. For further details and to review the guidelines, please refer to the following URL:

https://www.bristol.ac.uk/media-library/sites/alspac/documents/alspac-publications-checklist.pdf.

Specific details on the ethics committee and institutional review boards that approved aspects of the study can be found on the University of Bristol's webpage, please refer to the following URL:

https://bristol.ac.uk/alspac/researchers/research-ethics/

Supplementary Methods:

Classification of Variables:

1. **Initial Review and Categorization**: Each variable in the labelled dataset was manually reviewed to identify its inherent statistical and practical characteristics. This in-depth analysis allowed for a comprehensive classification into one of four primary types: ordinal, nominal, binary, and continuous.

2. **Alignment with PHESANT**: The manual classification approach was inspired by PHESANT's methodology [22]. The manual review ensured high accuracy, making the dataset robust for the study's objectives.

3. **Nominal to Binary Transformation**: All the nominal variables were further classified into binary variables. The decision to perform this additional classification was based on two primary factors:

- **Simplification for Analysis**: Converting nominal variables to binary forms allowed for more straightforward interpretability in the model, particularly when the nominal variables had only two substantive categories.

- **Alignment with Research Goals**: Binary variables clarified specific outcomes, making them more suited to the study's hypothesis-free approach.

By providing this level of detail in the classification process, our study offers a clear roadmap for those interested in replicating the analysis or applying similar methodologies in future research.

Data Cleaning and Pre-processing:

A. Classification of Variables:

Each variable was classified into nominal, ordinal, interval, or ratio scales based on its characteristics and role in the analysis. This classification allows for applying the most appropriate statistical tests and analytical techniques, thereby increasing the study's internal validity.

B. Handling Negative Values:

Negative values were assessed according to the context and nature of the variable in which they appeared. In variables where negative values were not logically permissible, these were corrected to ensure internal consistency and to prevent potential biases in the analysis.

C. Imputation of Missing Values:

The Random Forest algorithm used in this study is sensitive to missing values. As a result, a sophisticated imputation method was employed to fill in these gaps. Using imputation reduced the likelihood of biases arising from incomplete data, thus improving the robustness of the analyses.

D. Outlier Analysis:

The presence of outliers can affect the outcomes and reduce the validity of the findings. The methods were employed to identify and assess outliers in the data. The absence of substantial

outliers after this step served as an indicator of the reliability of the dataset, further confirming the rigour of the cleaning process.

E. Ensuring Data Quality:

Several quality checks were implemented to verify the data's internal consistency, accuracy, and completeness. These checks involved cross-referencing the data against source documents and performing statistical tests for internal consistency. The quality checks further substantiated the reliability of the data and the subsequent analyses.

By implementing these steps, the methodology demonstrated rigour and precision, enhancing the reliability and validity of the study's findings.

Statistical and Machine Learning Methods:

In the pre-processing phase, one critical aspect was identifying and removing zero-variance predictors [16, 17]. These predictors are characterised by the absence of variance across all observations, meaning all values are the same. As a result, they offered no valuable information to a predictive model and were subsequently removed. These can be systematically identified through statistical functions that compute the variance for each predictor, flagging those with zero variance for removal.

Additionally, the methodology considers near-zero variance predictors [8]. These are predictors with an extremely low level of variance and are nearly constant across the dataset. Despite not being perfectly constant like zero-variance predictors, near-zero variance predictors also contribute minimal utility to predictive modelling. Functions like nearZeroVar () in R were used to identify these by comparing the most frequent value to the second most frequent value and calculating the percentage of distinct values compared to the number of samples.

The steps to remove both zero and near-zero variance predictors are crucial for increasing the computational efficiency of the model and its ultimate accuracy. These actions are part of a rigorously planned data pre-processing strategy designed to ensure the reliability and robustness of subsequent analyses.

Our research methodology employs a 10-fold cross-validation technique integrated within the Random Forest algorithm to optimise the ROC curve, focusing specifically on maximising the

Area Under the Curve (AUC). Notably, the existing code does not delve into hyperparameter tuning practices like grid search, which leaves room for potential performance enhancements. Our cross-validation approach is termed 'detailed,' as it goes beyond conventional practices by saving final predictions and calculating class probabilities for each fold. These added metrics are summarised through the twoClassSummary function, making our evaluation particularly apt for binary classification problems such as dropout prediction. A 10-fold cross-validation was strategically made to balance computational efficiency with robust performance evaluation. This method ensures that each data point appears in the validation set precisely once, thus mitigating the risk of model overfitting or underfitting. While the central focus of our methodology lies in ROC and AUC metrics, we also acknowledge the relevance of additional performance indicators like accuracy, precision, and recall, especially in cases of class imbalance. Although these metrics are not explicitly calculated in the current code, they can be seamlessly integrated using the twoClassSummary function. Our methodology offers a transparent, robust, and multidimensional framework for model evaluation, addressing key questions and concerns raised during the review process.

In this research, the Gini decrease metric is crucial for pinpointing key predictors within the Random Forest models employed. Originating from the concept of Gini impurity, which quantifies the level of disorder or randomness in a dataset, the Gini decrease helps evaluate how individual predictors contribute to the overall model accuracy. The algorithm calculates the initial Gini impurity at each node when constructing each decision tree within the Random Forest based on the existing class distribution. Subsequently, the model evaluates potential splits by computing the weighted Gini impurity for the resulting child nodes.

The Gini decrease for each split is determined by the reduction in impurity from the parent node to the child nodes, weighted by the proportion of samples that move to each child node. The predictor offering the most significant Gini decrease is then selected to perform the split. This process is carried out across multiple trees to obtain an aggregated measure of each predictor's contribution to decreasing the Gini impurity.

The predictors are subsequently ranked based on their average Gini decrease over all the decision trees in the Random Forest. A higher average Gini decrease indicates greater importance, revealing that the predictor is pivotal in enhancing the model's classification performance. This metric not only adds a layer of interpretability to our model but also guides future data collection and feature engineering by highlighting predictors of significance.

Therefore, utilising the Gini decrease metric is vital for the rigour and coherence of this study, offering insights into the factors most influential in affecting dropout rates in both clinic and questionnaire settings.